

The controlled thermodynamic integral for Bayesian model comparison

Chris J. Oates, Theodore Papamarkou, Mark Girolami

Abstract

Bayesian model comparison relies upon the model evidence, yet for many models of interest the model evidence is unavailable in closed form and must be approximated. Many of the estimators for evidence that have been proposed in the Monte Carlo literature suffer from high variability. This paper considers the reduction of variance that can be achieved by exploiting control variates in this setting. Our methodology is based on thermodynamic integration and applies whenever the gradient of both the log-likelihood and the log-prior with respect to the parameters can be efficiently evaluated. Results obtained on regression models and popular benchmark datasets demonstrate a significant and sometimes dramatic reduction in estimator variance and provide insight into the wider applicability of control variates to Bayesian model comparison.

Keywords. model evidence, control variates, variance reduction

Author Footnote. Chris J. Oates (E-mail: c.oates@warwick.ac.uk) is Research Fellow, Theodore Papamarkou (E-mail: t.papamarkou@warwick.ac.uk) is Research Associate and Mark Girolami (E-mail: m.girolami@warwick.ac.uk) is Professor, Department of Statistics, University of Warwick, Coventry, CV4 7AL. This work was supported by UK EPSRC EP/D002060/1, EP/J016934/1, EU Grant 259348 (Analysing and Striking the Sensitivities of Embryonal Tumours) and a Royal Society Wolfson Research Merit Award. The authors are grateful to Christian Robert for discussions on the use of control variates in this setting.

1 Introduction

In hypothesis-driven research we are presented with data \mathbf{y} that is assumed to have arisen under one of two (or more) putative models m_i characterised by a probability density $p(\mathbf{y}|m_i)$. Given *a priori* model probabilities $p(m_i)$, the data \mathbf{y} induce *a posteriori* probabilities $p(m_i|\mathbf{y})$ that are the basis for Bayesian model comparison. Since any prior probability distribution gets transformed to a posterior probability distribution through consideration of the data, the transformation itself represents the evidence

provided by the data (Kass and Raftery, 1995). For the simple case of two models, this transformation follows from Bayes' rule as

$$\underbrace{\frac{p(m_2|\mathbf{y})}{p(m_1|\mathbf{y})}}_{\text{posterior odds}} = \underbrace{\frac{p(\mathbf{y}|m_2)}{p(\mathbf{y}|m_1)}}_{\text{Bayes factor } B_{21}} \times \underbrace{\frac{p(m_2)}{p(m_1)}}_{\text{prior odds}}. \quad (1)$$

Thus the influence of the data on the posterior probability distribution is captured through that Bayes factor B_{21} in favour of Model 2 over Model 1. Rearranging, we can interpret the Bayes factor as the ratio of the posterior odds to the prior odds. A natural approach to computation of Bayes factors is to directly compute the evidence

$$p(\mathbf{y}|m_i) = \int p(\mathbf{y}|\boldsymbol{\theta}, m_i)p(\boldsymbol{\theta}|m_i)d\boldsymbol{\theta}, \quad (2)$$

provided by data \mathbf{y} in favour of model m_i , where $\boldsymbol{\theta}$ are parameters associated with model m_i . Yet for almost all models of interest, the evidence is unavailable in closed form and must be approximated. Numerous techniques have been proposed to approximate the model evidence (Eqn. 2), a selection of which includes *path sampling* (Ogata, 1989; Gelman and Meng, 1998), *harmonic means* (Gelfand and Dey, 1994), *Chib's method* (Chib and Jeliazkov, 2001), *nested sampling* (Skilling, 2006), *particle filters* (Del Moral *et al.*, 2006), *multicanonical algorithms* (Marinari and Parisi, 1992; Geyer and Thompson, 1995), *approximate Bayesian computation* (Didelot *et al.*, 2011) and *variational approximations* (Corduneanu and Bishop, 2001). Alternatively one can directly target the Bayes factor B_{21} that compares between two models. Here too numerous methods have been proposed, including *importance sampling* (Gelman and Meng, 1998; Chen *et al.*, 2000), *ratio importance sampling* (Torrie and Valleau, 1977), *bridge sampling* (Gelman and Meng, 1998; Chen *et al.*, 2000), *sequential Monte Carlo* (Zhou *et al.*, 2013), *annealed importance sampling* (Neal, 2001), *reversible-jump Markov chain Monte Carlo* (MCMC; Green, 1995) and also again *approximate Bayesian computation* (Toni *et al.*, 2009). Recent reviews of these methodologies include Vyshemirsky

and Girolami (2008); Marin and Robert (2010); Friel and Wyse (2012).

Of the estimators of evidence that are based on Monte Carlo sampling, it remains the case that estimator variance can in general be extremely high. General approaches to reduction of Monte Carlo error that have been proposed in the literature include *antithetic variables* (Green and Han, 1992), *control variates* and *Rao-Blackwellisation* (Robert and Casella, 2004), *Riemann sums* (Philippe and Robert, 2001) and a plethora of MCMC schemes that aim to improve mixing (e.g. Girolami and Calderhead, 2011). These methods could all be used to reduce the variance of estimators for model evidence that are based on computing Monte Carlo expectations. In this paper we extend the *zero-variance* (ZV) control variate technique, introduced in the physics literature by Assaraf and Caffarel (1999), to estimators of model evidence that are based on MCMC and *thermodynamic integration* (TI; Frenkel and Smit, 2002). The methodology applies whenever the gradient of the log-likelihood (and the log-prior) can be evaluated and therefore can be used “for free” when differential geometric sampling schemes are employed in construction of the Markov chain (Papamarkou *et al.*, 2014). Theoretical results are provided that guide maximal variance reduction in practice. Results on popular benchmark datasets demonstrate a substantial reduction in variance compared to existing estimators and the method is shown to be exact in the special case of Bayesian linear regression.

The paper proceeds as follows: Section 2 recalls key ideas from TI and ZV that we use in our methodology. In section 3 we derive control variates for TI and provide theoretical results that guide maximal variance reduction in practice. Section 4 compares the proposed methodology to the state-of-the-art estimators of model evidence applied to popular benchmark datasets. Section 5 investigates scenarios where the proposed methodology is likely to fail. Finally section 6 provides more general insight into the use of control variates in estimation of model evidence, drawing an important distinction between “equilibrium” and “non-equilibrium” estimators that determines whether or not control variates may be applicable.

2 Background

2.1 Thermodynamic integration

Path sampling and the closely related technique of TI emerged from the physics community as a computational approach to compute normalising constants (Gelman and Meng, 1998). Recent empirical investigations, including Vyshemirsky and Girolami (2008); Friel and Wyse (2012), have revealed that TI is among the most promising approach to estimation of model evidence. Below we provide relevant background on TI, referring the reader to Calderhead and Girolami (2009) for a detailed discussion of implementational details.

TI targets the model evidence directly; in what follows we therefore implicitly condition upon a model m and aim to compute the evidence $p(\mathbf{y}) = p(\mathbf{y}|m)$ provided by data \mathbf{y} in favour of model m . Following the presentation of Friel and Pettitt (2008), the *power posterior* is defined as $p(\boldsymbol{\theta}|\mathbf{y}, t) = p(\mathbf{y}|\boldsymbol{\theta})^t p(\boldsymbol{\theta}) / \mathcal{Z}_t(\mathbf{y})$ where the normalising constant is given by $\mathcal{Z}_t(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})^t p(\boldsymbol{\theta}) d\boldsymbol{\theta}$. Here t is known as an *inverse temperature* parameter and by analogy the process of increasing t is known as *annealing*. Note that $p(\boldsymbol{\theta}|\mathbf{y}, t = 0)$ is the density of the prior distribution, whereas $p(\boldsymbol{\theta}|\mathbf{y}, t = 1)$ is the density $p(\boldsymbol{\theta}|\mathbf{y})$ of the posterior distribution. Varying $t \in (0, 1)$ produces a continuous path between these two distributions and in this paper it is assumed that all intermediate distributions exist and are well-defined. The normalising constant $\mathcal{Z}_0(\mathbf{y})$ is equal to one and $\mathcal{Z}_1(\mathbf{y})$ is equal to $p(\mathbf{y})$, the model evidence that we aim to estimate.

The standard thermodynamic identity is

$$\log(p(\mathbf{y})) = \int_0^1 \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}, t} \log(p(\mathbf{y}|\boldsymbol{\theta})) dt \quad (3)$$

where the expectation in the integrand is with respect to the power posterior whose density is given above. The correctness of Eqn. 3 is established in e.g. Friel and Pettitt (2008). In TI, this one-dimensional integral is evaluated numerically using a

quadrature approximation over a discrete temperature ladder, whereas in the related approach of path sampling this integral is evaluated using MCMC. Note that the use of quadrature methods introduces bias into the estimator of model (log-)evidence; it is therefore important to select an accurate quadrature approximation (Appendix A).

2.2 Control variates and the ZV technique

Control variates are often employed when we aim to estimate, with reduced variance, the expectation $\mathbb{E}_\pi[g(\boldsymbol{\theta})]$ of a function $g(\boldsymbol{\theta})$ of a random variable $\boldsymbol{\theta}$ that is distributed according to a (possibly unnormalised) density $\pi(\boldsymbol{\theta})$. In this paper we focus on real-valued $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$ and we aim to approximate

$$\mathbb{E}_\pi[g(\boldsymbol{\theta})] = \frac{\int g(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int \pi(\boldsymbol{\theta})d\boldsymbol{\theta}}. \quad (4)$$

The generic control variate principle relies on constructing an auxiliary function $\tilde{g}(\boldsymbol{\theta}) = g(\boldsymbol{\theta}) + h(\boldsymbol{\theta})$ that satisfies $\mathbb{E}_\pi[h(\boldsymbol{\theta})] = 0$ and so $\mathbb{E}_\pi[\tilde{g}(\boldsymbol{\theta})] = \mathbb{E}_\pi[g(\boldsymbol{\theta})]$. Write $\mathbb{V}_\pi[g(\boldsymbol{\theta})]$ for the variance of the function $g(\boldsymbol{\theta})$ of a random variable $\boldsymbol{\theta}$ whose (unnormalised) density is $\pi(\boldsymbol{\theta})$. In many cases it is possible to choose $h(\boldsymbol{\theta})$ such that $\mathbb{V}_\pi[\tilde{g}(\boldsymbol{\theta})] < \mathbb{V}_\pi[g(\boldsymbol{\theta})]$, leading to a reduction in Monte Carlo variance. Intuitively, greater variance reduction can occur when $h(\boldsymbol{\theta})$ is negatively correlated with $g(\boldsymbol{\theta})$ under $\pi(\boldsymbol{\theta})$, since much of the randomness “cancels out” in the auxiliary function $\tilde{g}(\boldsymbol{\theta})$. In classical literature $h(\boldsymbol{\theta})$ is formed as a sum $\phi_1 h_1(\boldsymbol{\theta}) + \dots \phi_m h_m(\boldsymbol{\theta})$ where the $h_i(\boldsymbol{\theta})$ have zero mean under $\pi(\boldsymbol{\theta})$ and are known as *control variates*, whilst ϕ_i are coefficients that must be specified. For estimation based on Markov chains, Andradóttir *et al.* (1993) proposed control variates for discrete state spaces. Later Mira *et al.* (2003) extended this approach to continuous state spaces, observing that the optimal choice of $h(\boldsymbol{\theta})$ is intimately associated with the solution of the Poisson equation $h(\boldsymbol{\theta}) = \mathbb{E}_\pi[g(\boldsymbol{\theta})] - g(\boldsymbol{\theta})$ and proposing to solve this equation numerically. Further work on constructing control variates for Markov chains includes Hammer and Tjelmeland (2008) for Metropolis-Hastings chains and

Dellaportas and Kontoyiannis (2012) for Gibbs samplers.

In this paper we consider the particularly tractable class of ZV control variates that are expressed as functions of the gradient $\nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{\theta})$ of the log-target density (i.e. the score function). More specifically, Mira *et al.* (2013) proposed to use

$$h(\boldsymbol{\theta}) = -\frac{1}{2}\Delta_{\boldsymbol{\theta}}[P(\boldsymbol{\theta})] + \nabla_{\boldsymbol{\theta}}[P(\boldsymbol{\theta})] \cdot \mathbf{z}(\boldsymbol{\theta}) \quad (5)$$

where the *trial function* $P(\boldsymbol{\theta})$ is a polynomial in $\boldsymbol{\theta}$ and

$$\mathbf{z}(\boldsymbol{\theta}) = -\frac{1}{2}\nabla_{\boldsymbol{\theta}}[\log(\pi(\boldsymbol{\theta}))] \quad (6)$$

is proportional to the score function. In this paper we adopt the convention that both $\boldsymbol{\theta}$ and $\mathbf{z}(\boldsymbol{\theta})$ are $d \times 1$ vectors. The thermodynamic identity (Eqn. 3) is based on expected values of log-likelihoods $\log(\pi(\boldsymbol{\theta}))$. Since $\mathbf{z}(\boldsymbol{\theta})$ is closely related to $\log(\pi(\boldsymbol{\theta}))$, ZV control variates appear as a natural strategy to achieve variance reduction in TI. As shown in Mira *et al.* (2013), ZV control variates arise naturally in certain Gaussian models, leading, in some cases, to exact (i.e. deterministic) estimators that have zero variance. Intuitively, any density $\pi(\boldsymbol{\theta})$ that approximates a Gaussian forms a suitable candidate for implementing the ZV scheme. Theoretical conditions for asymptotic unbiasedness of ZV have been established (Appendix B).

ZV control variates are particularly tractable for two reasons: (i) For many models of interest it is possible to obtain a closed-form expression for Eqn. 5, compared to alternatives that require numerical solution of the Poisson equation; (ii) As recently noticed by Papamarkou *et al.* (2014), the ZV technique can be applied essentially “for free” inside differential-geometric MCMC sampling schemes for which the score function is a pre-requisite for sampling (Girolami and Calderhead, 2011).

3 Methodology

In section 3.1 we develop a control variate scheme for the estimation of model evidence, taking TI as our base estimator whose variance we propose to reduce. The main methodological challenge in this setting is the elicitation of both the optimal control variate coefficients ϕ and the optimal temperature ladder that underlies TI. In section 3.2 we derive optimal expressions for both these quantities and in section 3.3 we describe how coefficients and temperature ladders are selected in practice.

3.1 The controlled thermodynamic integral

Taking the target density $\pi(\theta)$ to be the power posterior $p(\theta|\mathbf{y}, t)$, it follows from Eqn. 6 that

$$\mathbf{z}(\theta|\mathbf{y}, t) = -\frac{t}{2} \frac{\nabla_{\theta} p(\mathbf{y}|\theta)}{p(\mathbf{y}|\theta)} - \frac{1}{2} \frac{\nabla_{\theta} p(\theta)}{p(\theta)}. \quad (7)$$

The ZV control variates (Eqn. 5) are then

$$h(\theta|\mathbf{y}, t) = -\frac{1}{2} \Delta_{\theta}[P(\theta|\phi(\mathbf{y}, t))] + \nabla_{\theta}[P(\theta|\phi(\mathbf{y}, t))] \cdot \mathbf{z}(\theta|\mathbf{y}, t) \quad (8)$$

where $\mathbf{z}(\theta|\mathbf{y}, t)$ is as defined in Eqn. 7. Here the coefficients $\phi \equiv \phi(\mathbf{y}, t)$ of the polynomial P will in general depend on both the data \mathbf{y} and inverse temperature t . Integrating these control variates into TI, we obtain the “controlled thermodynamic integral” (CTI)

$$\log(p(\mathbf{y})) = \int_0^1 \mathbb{E}_{\theta|\mathbf{y}, t}[\log(p(\mathbf{y}|\theta)) + h(\theta|\mathbf{y}, t)] dt. \quad (9)$$

In order to use CTI to estimate the model (log-)evidence we need to specify both (i) polynomial coefficients $\phi(\mathbf{y}, t)$ and (ii) an appropriate discretisation $0 = t_0 < t_1 < \dots < t_m = 1$ (the *temperature ladder*) of the one dimensional integral. Specification of

both polynomial coefficients and temperature ladder should be targeted at minimising the variance of CTI (see below).

3.2 Optimal coefficients and ladders

We derive the jointly optimal, variance-minimising, polynomial coefficients and temperature ladder. For the latter, note that there is a surjective mapping from partitions $0 = t_0 < t_1 < \dots < t_m = 1$ to probability distributions on $[0, 1]$ with density function $p(t)$ that is given by $\int_0^{t_i} p(s)ds = \frac{i}{m}$. For the development below it is convenient to focus on optimising the density $p(t)$, mapping back to the temperature ladder during implementation (see section 3.3 below). For clarity of the exposition write $g(\boldsymbol{\theta}) = \log(p(\mathbf{y}|\boldsymbol{\theta}))$ where we temporarily suppress dependence on both data \mathbf{y} and model m . The CTI identity can be rewritten as

$$\log(p(\mathbf{y})) = \int_0^1 \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y},t}[g(\boldsymbol{\theta}) + h(\boldsymbol{\theta}|t)]dt = \mathbb{E}_{\boldsymbol{\theta},t|\mathbf{y}} \left[\frac{g(\boldsymbol{\theta}) + h(\boldsymbol{\theta}|t)}{p(t)} \right] \quad (10)$$

where the final expectation is taken with respect to the distribution with density $p(\boldsymbol{\theta}, t|\mathbf{y}) = p(\boldsymbol{\theta}|\mathbf{y}, t)p(t)$. Under an approximation that Monte Carlo samples are obtained independently, so-called “perfect transitions”, the variance of the estimator of model (log-)evidence is given by

$$\frac{1}{N} \left\{ \int_0^1 \frac{\mathbb{E}_{\boldsymbol{\theta}|\mathbf{y},t}[(g(\boldsymbol{\theta}) + h(\boldsymbol{\theta}|t))^2]}{p(t)} dt - [\log(p(\mathbf{y}))]^2 \right\} \quad (11)$$

where N is the number of Monte Carlo samples.

The optimal choice of polynomial coefficients $\phi(t)$ and temperature ladder $p(t)$ are defined as the pair that jointly minimise Eqn. 11. Specifically, we seek to minimise the Lagrangian

$$\int_0^1 \frac{\mathbb{E}_{\boldsymbol{\theta}|\mathbf{y},t}[(g(\boldsymbol{\theta}) + h(\boldsymbol{\theta}|t))^2]}{p(t)} dt + \lambda \int_0^1 p(t) \quad (12)$$

over $(p, \phi) : [0, 1] \rightarrow \mathbb{R}^{e+1}$ where e is the dimension of ϕ and depends on the degree of the polynomial $P(\theta|\phi)$ that is being employed. Here λ is a Lagrange multiplier that will be used to ensure $\int p(t)dt = 1$. Below we consider degree 1 polynomials $P(\theta|\phi) = \theta^T \phi$ so that $h(\theta|t) = \phi(t)^T \mathbf{z}(\theta|t)$ but the derivation applies analogously to higher degree polynomials, as explained in Appendix C. The solution (p^*, ϕ^*) of the Lagrangian optimisation problem (Eqn. 12) is

$$\phi^*(t) = -\mathbb{V}_{\theta|y,t}^{-1}[\mathbf{z}(\theta)]\mathbb{E}_{\theta|y,t}[g(\theta)\mathbf{z}(\theta)] \quad (13)$$

$$p^*(t) \propto \sqrt{\mathbb{E}_{\theta|y,t}[g(\theta)^2] - \mathbb{E}_{\theta|y,t}[g(\theta)\mathbf{z}(\theta)]^T \mathbb{V}_{\theta|y,t}[\mathbf{z}(\theta)]^{-1} \mathbb{E}_{\theta|y,t}[g(\theta)\mathbf{z}(\theta)]} \quad (14)$$

where $\mathbb{V}_{\theta|y,t}[\mathbf{z}(\theta)]$ and $\mathbb{E}_{\theta|y,t}[g(\theta)\mathbf{z}(\theta)]$ denote respectively variance and cross-covariance matrices (since $\mathbb{E}_{\theta|y,t}[\mathbf{z}(\theta)] = \mathbf{0}$). Notice that the optimal temperature ladder for CTI is not the same as the optimal ladder for standard TI, which is given by $p^*(t) \propto \sqrt{\mathbb{E}_{\theta|y,t}[g(\theta)^2]}$ (Calderhead and Girolami, 2009).

It can be shown (Rubinstein and Marcus, 1985) that this choice of polynomial coefficients $\phi = \phi^*$ is characterised as the minimiser of the variance ratio

$$R(t) := \frac{\mathbb{V}_{\theta|y,t}[g(\theta) + \phi(t)^T \mathbf{z}(\theta|t)]}{\mathbb{V}_{\theta|y,t}[g(\theta)]} \quad (15)$$

and at this minimum

$$R(t) = 1 - \text{Corr}_{\theta|y,t}[g(\theta), \phi^T \mathbf{z}(\theta)]^2, \quad (16)$$

so that greater variance reduction is expected in the case where a linear combination of the elements of the vector $\mathbf{z}(\theta)$ is highly correlated with the target function $g(\theta)$.

3.3 Implementation

For most models of interest both Eqn. 13 and Eqn. 14 do not possess closed-form expressions and it becomes necessary to employ estimates or approximations to the optimal values. We begin by noting that Eqn. 13 actually defines the optimal, variance-minimising, coefficients independently of the choice of temperature ladder $p(t)$; this is directly verified from the Euler-Lagrange equations applied to $\phi : [0, 1] \rightarrow \mathbb{R}^e$ where $p(t)$ is held fixed. This observation allows us to discuss these two aspects of the implementation separately:

3.3.1 Polynomial coefficients

Optimal coefficients for control variates are typically estimated based on the same sequence of MCMC samples that will subsequently be used to compute the controlled expectations (Robert and Casella, 2004). Specifically, to estimate the optimal control variate coefficients $\phi^*(t)$ we exploit MCMC samples to estimate both the covariance $\hat{\mathbb{V}}_{\theta|y,t}[\mathbf{z}(\theta)]$ and the cross-covariance $\hat{\mathbb{E}}_{\theta|y,t}[g(\theta)\mathbf{z}(\theta)]$. These estimates are then plugged directly into Eqn. 13 in order to obtain an estimate

$$\phi^*(t) \approx -\hat{\mathbb{V}}_{\theta|y,t}[\mathbf{z}(\theta)]^{-1}\hat{\mathbb{E}}_{\theta|y,t}[g(\theta)\mathbf{z}(\theta)] \quad (17)$$

for the optimal coefficients. Further discussion of “plug-in” estimators for control coefficients can be found in Dellaportas and Kontoyiannis (2012).

3.3.2 Temperature ladder

For estimating the optimal temperature ladder of Eqn. 14, one obvious numerical approach would be to firstly estimate $p^*(t)$ up to proportionality over a uniform grid $\{t_i\}$, using a preliminary MCMC run to estimate both $\mathbb{E}_{\theta|y,t}[g(\theta)^2]$ and the covariance and cross-covariance matrices $\mathbb{V}_{\theta|y,t}[\mathbf{z}(\theta)]$ and $\mathbb{E}_{\theta|y,t}[g(\theta)\mathbf{z}(\theta)]$. Then nonparametric density estimation could be applied in order to obtain an estimate for the optimal

ladder $\{t_i\}$. However this two-step procedure is computationally burdensome. Neal (1996) showed that a geometric temperature ladder is optimal for annealing on the scale parameter of a Gaussian and Behrens *et al.* (2012) extended this result to target distributions of the same form as $g(\boldsymbol{\theta})$, which includes Gaussians. In this paper we fix a quintic temperature ladder $t_i = (i/50)^5$ for use in all applications; this ladder is widely used in the TI literature and has demonstrated strong performance in empirical studies (e.g. Calderhead and Girolami, 2009; Friel *et al.*, 2014). The question of how to select appropriate temperature ladders in practice is an ongoing area of research and the recent contributions of Miasojedow *et al.* (2012); Behrens *et al.* (2012); Zhou *et al.* (2013); Friel *et al.* (2014) are compatible with our methodology.

3.3.3 Quadrature

The second order quadrature method of Friel *et al.* (2014), described in Appendix A, requires us also to estimate the variance $\mathbb{V}_{\boldsymbol{\theta}|\mathbf{y},t}[\log(p(\mathbf{y}|\boldsymbol{\theta}))]$ at each step in the temperature ladder. In experiments below we use ZV control variates to estimate this variance, using the identity

$$\mathbb{V}_{\boldsymbol{\theta}|\mathbf{y},t}[\log(p(\mathbf{y}|\boldsymbol{\theta}))] = \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y},t} [\log(p(\mathbf{y}|\boldsymbol{\theta})) - \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y},t}[\log(p(\mathbf{y}|\boldsymbol{\theta}))]]^2 \quad (18)$$

and applying control variates in the estimation of each of these expectations.

4 Applications

We present several empirical studies that compare CTI to the state-of-the-art TI estimators of Friel *et al.* (2014). In all applications below we base estimation on the output of a population MCMC sampler (Jasra *et al.*, 2007) limited to N iterations at each of the 51 rungs of the temperature ladder (a total of $51 \times N$ evaluations of the likelihood function). In brief, the within-temperature proposal was provided by the

manifold Metropolis-adjusted Langevin algorithm (mMALA) of Girolami and Calderhead (2011), whilst the between-temperature proposal randomly chooses a pair of (inverse) temperatures t_i and t_j , proposing to swap their state vectors with probability given by the Metropolis-Hastings ratio (Calderhead and Girolami, 2009). To ensure fairness, the same samples were used as the basis for all estimators of model evidence, ensuring that all estimators require essentially the same amount of computation (since the score function is computed as a matter of course in mMALA). Moreover, to explore the statistical properties of the estimators themselves, we generated 100 independent realisations of the population MCMC and thus 100 realisations of each estimator. Full details are provided in the Supplement.

4.1 Bayesian linear regression

4.1.1 Known precision

We begin with an analytically tractable problem in Bayesian linear regression. The (log-)likelihood function is given by

$$\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (19)$$

where \mathbf{y} is $n \times 1$, \mathbf{X} is $n \times d$ and $\boldsymbol{\beta}$ is $d \times 1$. In simulations below we took $\sigma = 1$, $d = 3$, $\boldsymbol{\beta} = [0, 1, 2]$. The design matrix \mathbf{X} was populated with $n = 100$ rows by drawing each entry independently from the standard normal distribution and then data \mathbf{y} were generated from $N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_{n \times n})$; both \mathbf{X} and \mathbf{y} were then fixed for all experiments below. From the Bayesian perspective we take a conjugate prior $\boldsymbol{\beta} \sim N(\mathbf{0}, \zeta^2 \mathbf{I}_{d \times d})$ with $\zeta = 1$. In this section we assume σ is fixed and known, but we relax this assumption in the next section. Thus the unknown model parameters here are $\boldsymbol{\theta} = \boldsymbol{\beta} \in \mathbb{R}^d$ and we aim to compute the evidence $p(\mathbf{y})$ by marginalising over these parameters. This example is an ideal benchmark since it is permissible to obtain many relevant quantities

in closed form; see Appendix D.1 for full details.

Before applying CTI we are required to check that the sufficient conditions for the unbiasedness of ZV estimators are satisfied (see Appendix B). This amounts to noticing that the tails of the power posterior $p(\boldsymbol{\beta}|\mathbf{y}, t)$ decay exponentially in $\boldsymbol{\beta}$ (Appendix D.1). Using the plug-in estimates (Eqn. 17) we obtain estimates for the optimal coefficients $\boldsymbol{\phi}^*$, that are shown in SFig. 6. For degree 2 polynomials we see that the plug-in estimator is deterministic. Indeed, by direct calculation we see that $\mathbf{z}(\boldsymbol{\beta}|\mathbf{y}, t)$ is an invertible affine transformation of the parameter vector

$$\mathbf{z}(\boldsymbol{\beta}|\mathbf{y}, t) = -\frac{t}{2\sigma^2}\mathbf{X}^T\mathbf{y} + \frac{1}{2}\boldsymbol{\Sigma}(t)^{-1}\boldsymbol{\beta} \quad (20)$$

where $\boldsymbol{\Sigma}(t) = (\frac{t}{\sigma^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\zeta^2}\mathbf{I})^{-1}$. This allows us to intuit that CTI based on degree 2 polynomials will produce an *exact* estimate of the (log-)evidence (up to quadrature error), as we explain below. Indeed, by another invertible affine transformation we can map $\mathbf{z}(\boldsymbol{\beta}|\mathbf{y}, t) \mapsto \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ which, when multiplied by the polynomial $P(\boldsymbol{\beta}|\boldsymbol{\phi}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T$ produces a quantity $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ that is perfectly correlated with the log-likelihood under the power posterior. It then follows from Eqn. 15 that CTI will possess zero variance. This argument is made rigorous in the Supplement.

In SFig. 7 we plot 100 independent estimates of the integrand $\mathbb{E}_{\boldsymbol{\beta}|\mathbf{y}, t}[g(\boldsymbol{\beta})]$ at each of the 51 temperatures in the ladder for polynomial trial functions of degree 0 (i.e. standard TI), 1 and 2. It is apparent that estimator variance is greatest at lower values of t ; this motivates the heavily skewed temperature ladder used by ourselves and others, as we wish to target our computational effort on this high-variance region. We quantify the reduction in estimator variance at an (inverse) temperature t using the variance ratio $R(t)$ as estimated from the MCMC samples. Fig. 1 shows that degree 1 polynomials achieve (on average) variance reduction at all temperatures, with the greatest reduction occurring in the region where t is small. This is encouraging as the region where t is small is most important for variance reduction of TI, as discussed

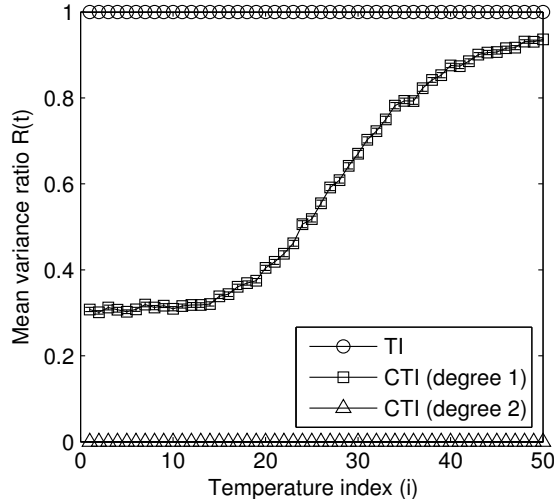


Figure 1: Bayesian linear regression, known precision. [Here we plot the mean variance ratio $R(t)$ computed over 100 independent runs of population MCMC using $N = 1000$ samples. Error bars show standard error of these mean estimates. The x-axis records the index i corresponding to (inverse) temperature $t_i = (i/50)^5$.]

above. For degree 2 polynomials we have $R(t) = 0$ for all t , which recapitulates the exactness of the CTI estimator in this example.

Finally we explore the quality of the estimators of model evidence themselves. For this model the (log-)evidence is available in closed form (Appendix D.1) and this allows us to compute the mean square error (MSE) over all 100 independent realisations of each estimator. Results, shown in Table 1, demonstrate that CTI with degree 2 polynomials achieves a 2-fold reduction in MSE compared to standard TI when both estimators are based on first order quadrature. However, first order quadrature is known to lead to significant estimator bias (Friel *et al.*, 2014) and when estimators are based instead on more accurate second order quadrature, CTI is seen to be approximately $10,000\times$ better than TI in terms of MSE; a dramatic difference. We also compared TI approaches against annealed importance sampling (AIS; Neal, 2001), as described in the Supplement. In this case CTI (degree 2) is over $10,000\times$ more accurate compared to AIS (SFig. 9a).

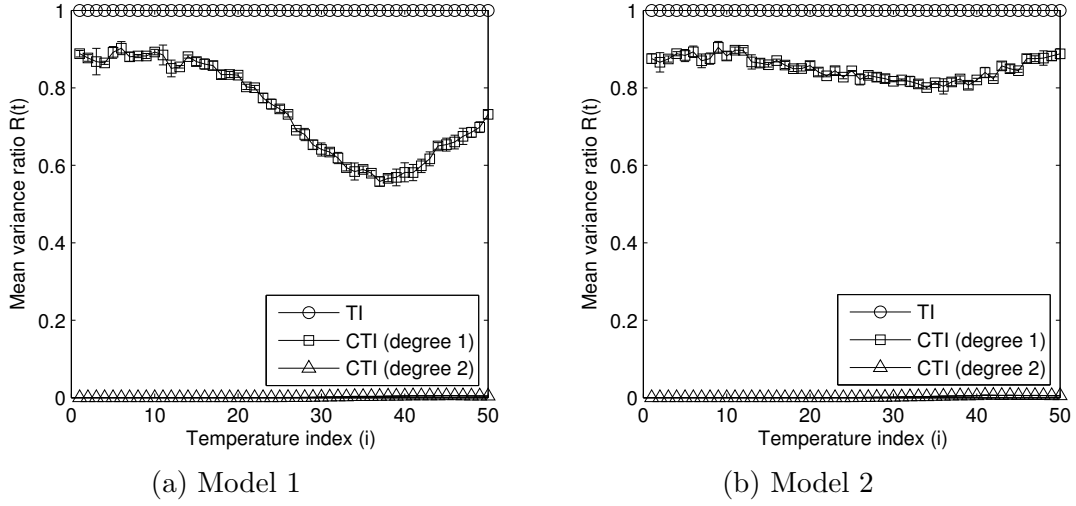


Figure 2: Bayesian linear regression, unknown precision. [Here we plot the mean variance ratio $R(t)$ computed over 100 independent runs of population MCMC using $N = 1000$ samples. Error bars show standard error of these mean estimates. The x-axis records the index i corresponding to (inverse) temperature $t_i = (i/50)^5$.]

4.1.2 Unknown precision (Radiata Pine)

We now relax the assumption of known precision $\tau = 1/\sigma^2$; we will see that in these circumstances CTI is no longer exact. Specifically we consider data from Williams (1959) on $n = 42$ specimens of *radiata* pine. This dataset is well known in the multivariate statistics literature and was recently used by Friel and Wyse (2012); Friel *et al.* (2014) in order to benchmark estimators of model evidence. Data consist of the maximum compression strength parallel to the grain y_i as a function of density x_i and density adjusted for resin content z_i . It is wished to determine whether the density or resin-adjusted density is a better predictor of compression strength parallel to the grain. Following Friel *et al.* (2014) we consider Bayesian model comparison between a pair of competing models:

$$\text{Model 1: } y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i, \quad \epsilon_i \sim N(0, \tau^{-1}) \quad (21)$$

$$\text{Model 2: } y_i = \gamma + \delta(z_i - \bar{z}) + \eta_i, \quad \eta_i \sim N(0, \lambda^{-1}) \quad (22)$$

Here \bar{x} and \bar{z} are the sample means of the x_i and z_i respectively. The priors for (α, β) and (γ, δ) are both Gaussian with common mean $\mathbf{B}_0 = [3000, 185]^T$ and precisions $\tau \mathbf{Q}_0, \lambda \mathbf{Q}_0$ where $\mathbf{Q}_0 = \text{diag}(0.06, 6)$. Both τ and λ were assigned gamma priors with shape 6 and rate 4×300^2 . To compare between these models we consider estimates for the log-Bayes factor $\log(B_{21})$ that are obtained as the difference between independent estimates for the log-evidence of each model.

This application is interesting for two reasons: Firstly, one can directly calculate the Bayes factor for this example as $B_{21} = 8.7086$, so that we have a gold standard performance benchmark. Secondly, when the precision τ (or λ) is unknown, ZV methods are no longer exact. We therefore have an opportunity to assess the performance of CTI in a non-trivial setting.

Formulae in Appendix D.2 demonstrate that the sufficient condition for unbiasedness of ZV methods is satisfied. Results in Fig. 2 show that CTI (degree 1) achieves a modest reduction in variance across temperatures t , whereas CTI (degree 2) achieves a massive variance reduction. Computing the MSE relative to the true Bayes factor we see that CTI (degree 2) is over $500\times$ more accurate compared to TI, though the variance of the estimator is not identically equal to zero in this case (Table 1). As before, MSE is further reduced as a result of applying second order quadrature. AIS performed slightly worse than the methods based on TI in this example (SFig. 9b).

4.2 Bayesian logistic regression (Pima Indians)

Here we examine data that contains instances of diabetes and a range of possible diabetes indicators for $n = 532$ women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona. This dataset is frequently used as a benchmark for supervised learning methods (e.g. Marin and Robert, 2010). Friel *et al.* (2014) considered seven predictors of diabetes recorded for this group; number of pregnancies (NP); plasma glucose concentration (PGC); diastolic blood pressure (BP); triceps skin

Precision	N	$\deg(P)$	Quadr.	M.S.E.	S.E.
(a) Known	1e3	0	1	2.3e-2	2.9e-3
			2	2.1e-2	2.5e-3
		1	1	1.8e-2	2.1e-3
			2	2.0e-2	2.5e-3
		2	1	1.2e-2	0
			2	2.2e-6	1.6e-7
	5e3	0	1	5.2e-3	7.2e-4
			2	4.0e-3	5.9e-4
		1	1	3.8e-3	6.0e-4
			2	3.4e-3	5.3e-4
		2	1	1.2e-3	0
			2	2.0e-7	2.7e-8
(b) Unknown	1e3	0	1	7.9e-3	1.2e-3
			2	7.7e-3	1.1e-3
		1	1	7.8e-3	1.0e-3
			2	7.6e-3	1.0e-3
		2	1	1.4e-5	2.0e-6
			2	1.3e-5	1.6e-6
	5e3	0	1	1.4e-3	1.8e-4
			2	1.3e-3	1.8e-4
		1	1	1.4e-3	2.0e-4
			2	1.4e-3	2.0e-4
		2	1	2.4e-6	3.0e-7
			2	1.5e-6	2.0e-7

Table 1: Bayesian linear regression with (a) known precision and (b) unknown precision. Mean square error (MSE) for estimates of the log-evidence in (a) and the Bayes factor in (b), based on 100 independent runs of population MCMC, along with estimates for standard errors (SE). $\dim(P)$ is the dimension of the ZV polynomial $P(\boldsymbol{\theta})$, with 0 denoting standard TI. Quadr. is the order of numerical quadrature scheme. N is the number of MCMC iterations.

fold thickness (TST); body mass index (BMI); diabetes pedigree function (DP) and age (AGE). Diabetes incidence y_i in person i is modelled by the binomial likelihood

$$p(\mathbf{y}|\boldsymbol{\beta}) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i}, \quad (23)$$

where the probability of incidence p_i for person i is related to the covariates $\mathbf{x}_{i,\bullet} = (1, x_{i,1}, \dots, x_{i,d})^T$ and the parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_d)$ by

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \mathbf{x}_{i,\bullet}\boldsymbol{\beta}. \quad (24)$$

Bayesian model comparison is desired to be performed between the two candidate models

$$\text{Model 1:} \quad \text{logit}(p) = \beta_0 + \beta_1 \text{NP} + \beta_2 \text{PGC} + \beta_3 \text{BMI} + \beta_4 \text{DP} \quad (25)$$

$$\text{Model 2:} \quad \text{logit}(p) = \beta_0 + \beta_1 \text{NP} + \beta_2 \text{PGC} + \beta_3 \text{BMI} + \beta_4 \text{DP} + \beta_5 \text{AGE} \quad (26)$$

subject to the prior belief $\boldsymbol{\beta} \sim N(\mathbf{0}, \tau^{-1} \mathbf{I})$. Following Friel and Wyse (2012) we set $\tau = 0.01$.

The unbiasedness criterion in Appendix B is seen to be satisfied and we have

$$\mathbf{z}(\boldsymbol{\beta}|\mathbf{y}, t) = -\frac{t}{2} \mathbf{X}^T (\mathbf{y} - \mathbf{p}) + \frac{\tau \boldsymbol{\beta}}{2} \quad (27)$$

where the i th row of \mathbf{X} is $\mathbf{x}_{i,\bullet}$. In Fig. 3 we see that degree 1 ZV methods achieve a greater variance reduction at smaller t , but moreover we see that degree 2 ZV methods continue to achieve a substantial variance reduction at all temperatures. In Table 2 we display the mean of each estimator \hat{B}_{21} , computed over all 100 independent runs of population MCMC, together with the standard deviation of this collection of estimates. We see that this variance reduction transfers to estimates of the Bayes factor themselves, where the standard deviation of the CTI estimators is approximately $20\times$ lower compared to estimators based on TI. Although no exact expression is available for B_{21} , Friel *et al.* (2014) computed the log-evidence for both models using an extended run of 2,000 temperatures and $N = 20,000$ iterations within standard TI. Their estimates were -257.2342 and -259.8519 respectively for Models 1 and 2, corresponding to an estimate of the Bayes factor of $B_{21} = -2.6177$. This estimate, obtained at considerable computational expense, closely matches the estimates obtained by CTI (degree 2), which is based on $800\times$ fewer evaluations of the likelihood function. AIS performs comparably with standard TI in this example (SFig. 9c).

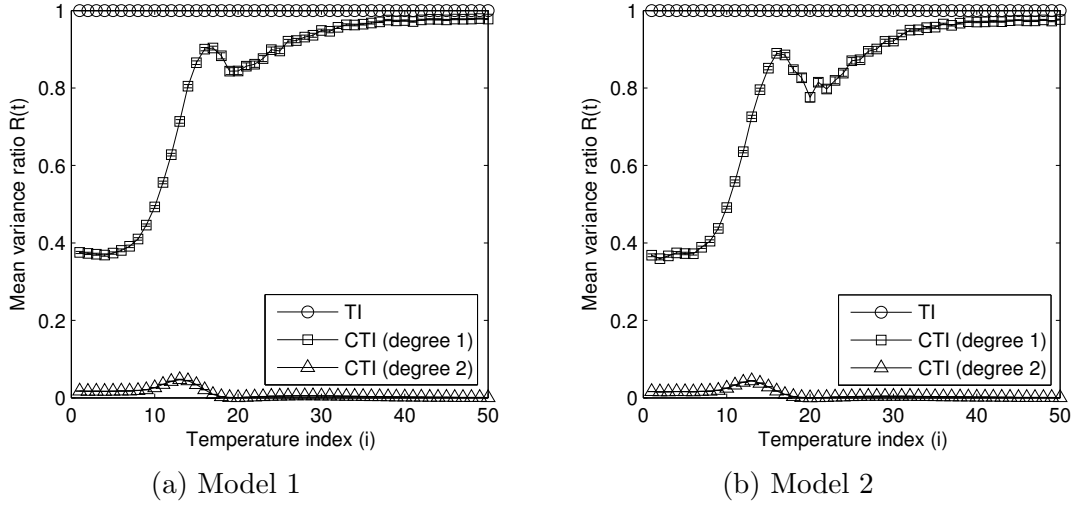


Figure 3: Bayesian logistic regression. [Here we plot the mean variance ratio $R(t)$ computed over 100 independent runs of population MCMC using $N = 1000$ samples. Error bars show standard error of these mean estimates. The x-axis records the index i corresponding to (inverse) temperature $t_i = (i/50)^5$.]

Model	N	$\deg(P)$	Quadr.	Mean B.F.	S.D.
(a) Logistic regression	1e3	0	1	-2.59	0.74
			2	-2.58	0.73
		1	1	-2.44	0.70
			2	-2.42	0.69
		2	1	-2.62	0.050
			2	-2.61	0.044
	5e3	0	1	-2.62	0.35
			2	-2.60	0.34
		1	1	-2.58	0.35
			2	-2.56	0.34
		2	1	-2.64	0.016
			2	-2.62	0.016
(b) Nonlinear ODEs	1e3	0	1	-3.75	0.31
			2	-3.74	0.31
		1	1	-3.69	0.31
			2	-3.69	0.31
		2	1	-3.57	0.27
			2	-3.56	0.27

Table 2: Estimates of the log-Bayes factor B_{21} , based on 100 independent runs of population MCMC. (a) Bayesian logistic regression. The actual Bayes factor, as computed by Friel *et al.* (2014), is $B_{21} = -2.6177$. (b) Nonlinear ODEs: Estimates of the log-Bayes factor B_{12} , based on 10 independent runs of population MCMC. $\dim(P)$ is the dimension of the ZV polynomial $P(\theta)$, with 0 denoting standard TI. Quadr. is the order of numerical quadrature scheme. N is the number of MCMC iterations. Mean BF and SD are the mean and standard deviation of the estimated Bayes factors.

5 Limitations of CTI

We have demonstrated, using standard benchmark datasets, that CTI is well-suited to Bayesian model comparison between regression models. Regression analyses continue to be widely applicable in disciplines such as econometrics, epidemiology, political science, psychology and sociology, so that these findings have significant implications. Nevertheless in many disciplines such as engineering, geophysics and systems biology, statistical models are significantly more complex, often based on a mechanistic understanding of the underlying process. Below we provide such an example and find that CTI offers little improvement over TI; this allows us to explore the limitations of our approach and, conversely, to understand in what circumstances it is likely to be successful.

5.1 A negative example (Goodwin Oscillator)

We consider nonlinear dynamical systems of the form

$$\frac{d\mathbf{x}}{ds} = \mathbf{f}(\mathbf{x}, s; \boldsymbol{\theta}), \quad \mathbf{x}(0) = \mathbf{x}_0. \quad (28)$$

We assume only a subset of the variables are observed under noise, so that $\mathbf{x} = [\mathbf{x}_a, \mathbf{x}_b]$ and \mathbf{y} is a d by n matrix of observations of the variables \mathbf{x}_a . Model comparison for systems specified by nonlinear differential equations is known to be profoundly challenging (Calderhead and Girolami, 2011).

Write $s_1 < s_2 < \dots < s_n$ for the times at which observations are obtained, such that $\mathbf{y}(s_j) = \mathbf{y}_{\bullet, j}$. We consider a Gaussian observation process with likelihood

$$p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{x}_0, \sigma) = \prod_{j=1}^n \mathcal{N}(\mathbf{y}(s_j) | \mathbf{x}_a(s_j; \boldsymbol{\theta}, \mathbf{x}_0), \sigma^2 \mathbf{I}) \quad (29)$$

where $\mathbf{x}_a(s_j; \boldsymbol{\theta}, \mathbf{x}_0)$ denotes the solution of the system in Eqn. 28. For the Gaussian

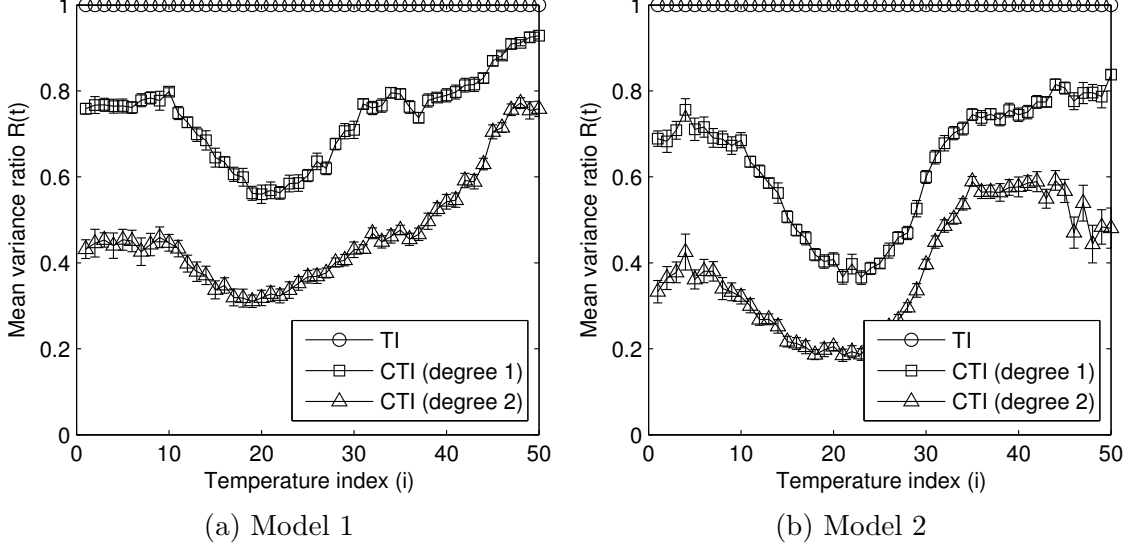


Figure 4: Nonlinear ODEs. [Here we plot the mean variance ratio $R(t)$ computed over 10 independent runs of population MCMC using $N = 1000$ samples. Error bars show standard error of these mean estimates. The x-axis records the index i corresponding to (inverse) temperature $t_i = (i/50)^5$.]

observation model it can be shown that a sufficient condition for unbiasedness of ZV is that the parameter prior density $p(\boldsymbol{\theta})$ vanishes faster than r^{d+k-2} when $r = \|\boldsymbol{\theta}\|_1 \rightarrow \infty$. Here $d = \dim \Theta$ and $k = 1$ is the degree of the polynomial that is being employed (see Appendix B).

Assuming the sufficient condition for ZV is satisfied, we have

$$z_i(\boldsymbol{\theta}) = -\frac{t}{2\sigma^2} \sum_{j=1}^n \mathbf{S}_{j,1:\dim \mathbf{x}_a}^i (\mathbf{y}(s_j) - \mathbf{x}_a(s_j; \boldsymbol{\theta}, \mathbf{x}_0)) - \frac{1}{2} \nabla_{\boldsymbol{\theta}} \log(p(\boldsymbol{\theta})) \quad (30)$$

where \mathbf{S}^i is a matrix of *sensitivities* with entries $S_{j,k}^i = \frac{\partial x_k}{\partial \theta_i}(s_j)$. Note that in Eqn. 30, $\mathbf{S}_{j,k}^k$ ranges over indices $1 \leq k \leq \dim \mathbf{x}_a$ corresponding only to the observed variables. In general the sensitivities \mathbf{S}^i will be unavailable in closed form, but may be computed numerically by augmenting the system of ordinary differential equations (ODEs) in Eqn. 28 as described in Appendix E. Indeed, these sensitivities are already computed when differential-geometric sampling schemes are employed, so that the evaluation of

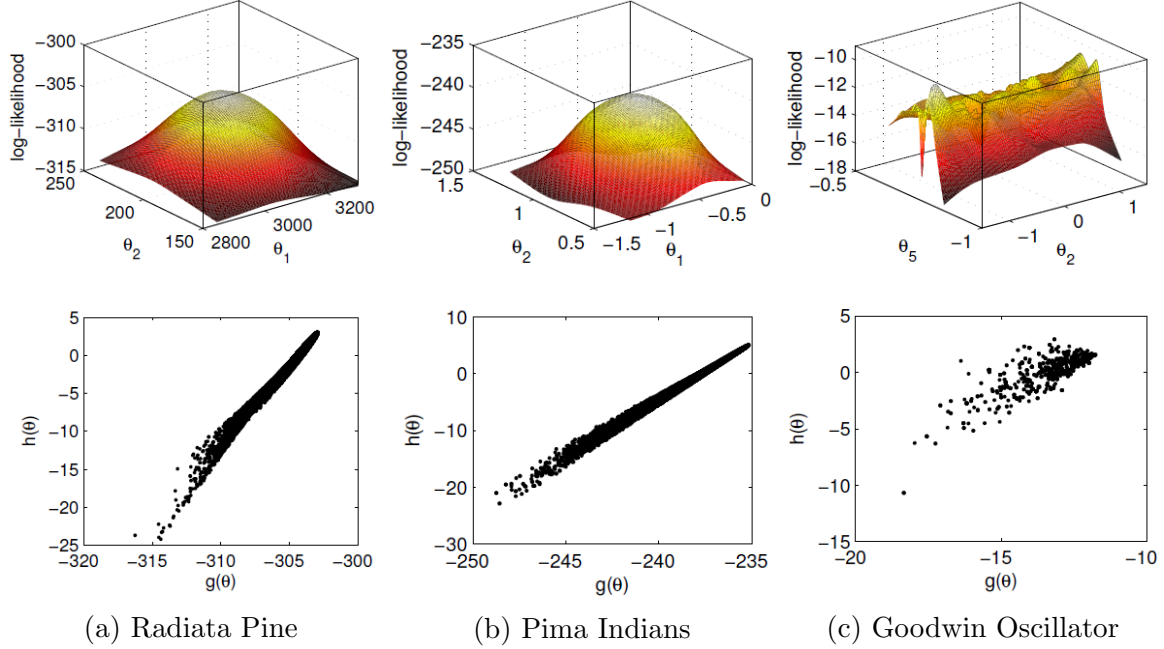


Figure 5: Comparing the likelihood surfaces and canonical correlations of different models. [Log-likelihood surfaces (top) for the (a) Radiata Pine and (b) Pima Indians examples can be well-approximated by a Gaussian and induce strong canonical correlation (bottom) between the (degree 2) control variates $h(\theta)$ and the log-likelihood $g(\theta)$ in the posterior. On the other hand, the log-likelihood surface for (c) Goodwin Oscillator is highly multi-modal and there is much weaker canonical correlation between the control variates and the log-likelihood.]

Eqn. 30 incurs negligible computational cost.

We focus on a dynamical model of oscillatory enzymatic control due to (Goodwin, 1965), that was recently considered in the context of Bayesian model comparison by Calderhead and Girolami (2009). This kinetic model, specified by a system of g ODEs, describes how a negative feedback loop between protein expression and mRNA transcription can induce oscillatory dynamics as experimentally observed in circadian regulation (Locke *et al.*, 2005). A full specification is provided in Appendix E. As shown in Calderhead and Girolami (2009), the *Goodwin oscillator* induces a highly multi-modal posterior distribution that renders estimation of the model evidence extremely challenging. We consider Bayesian comparison of two models; a simple model with one intermediate protein species ($g = 3$) and a more complex model with two

intermediate protein species ($g = 4$).

Fig. 4 demonstrates that in this extremely challenging example the benefits of control variate schemes that we have previously observed are heavily reduced. Since the variance ratio $R(t)$ is related to the canonical correlation between control variates and the log-likelihood under the power posterior (Eqn. 16), we hypothesise that the extreme multi-modality of the power posterior distribution is limiting the extent to which strong canonical correlation can be achieved. This is confirmed in Fig. 5 where we plot values of the target function $g(\theta)$ against the control variates $h(\theta)$ that are obtained from MCMC sampling in the posterior. We observe much reduced correlation in the case of the Goodwin oscillator that is a consequence of the complex nature of the likelihood surface. Turning to the Bayes factor itself, in Table 2 we display the mean of each estimator of the Bayes factor, together with the standard deviation of this collection of estimates. We find that CTI (degree 1) provides negligible reduction in variance and CTI (degree 2) provides an insignificant 15% reduction in variance.

In this example AIS consistently produced lower estimates for Bayes factors (SFig. 9d). This likely reflects the low number N of Monte Carlo iterations that are characteristic of such computationally demanding applications.

6 Discussion

To the best of our knowledge this is the first paper to consider the use of control variates for the purpose of Bayesian model comparison. Motivated by previous empirical studies, we focussed on TI estimators for the model (log-)evidence. However, in general, control variate techniques could be leveraged in Bayesian model comparison whenever estimators of the evidence (or Bayes factors) take the form of a Monte Carlo expectation. General control variate schemes for MCMC rely on the fact that the expectation of the control variates along the MCMC sample path will be approximately zero. We thus draw a distinction between “equilibrium” Monte Carlo estimators for the model

evidence, such as TI and path sampling, that require the underlying Markov chain to have converged, and “non-equilibrium” estimators such as AIS and sequential Monte Carlo that do not require convergence. The former class are amenable to existing control variate schemes whereas the latter are not. This motivates the “equilibration” of these non-equilibrium estimators.

Given its close connection with TI (Gelman and Meng, 1998), we considered whether an equilibrated version of AIS, that jointly samples from all rungs of the temperature ladder at once, would benefit from application of ZV control variates. In contrast to CTI, the controlled AIS estimator (CAIS) demonstrated an *increase* in variance compared to standard AIS. Full details are provided in the Supplement, in addition to results on each of the applications considered in this paper. To understand these counter-intuitive results, notice that control variates must be constructed simultaneously over all m rungs of the temperature ladder, so that for degree 1 polynomials we have to jointly estimate md coefficients, where d denotes the number of model parameters, and for degree 2 polynomials we have to jointly estimate $md(d+3)/2$ polynomial coefficients. To achieve this using the plug-in principle, we must estimate covariance matrices containing $\mathcal{O}(m^2d^2)$ and $\mathcal{O}(m^2d^4)$ entries respectively. Our results are therefore consistent with the finding that poor estimation of the polynomial coefficients can actually increase estimator variance (Glasserman, 2004). It remains unclear how to develop control variates for these non-equilibrium estimators.

We exploited the ZV control variate scheme due to Mira *et al.* (2013) that permits the automatic construction of control variates for any statistical model in which the gradient of the log-likelihood (and the log-prior) are available. More generally, we envisage that for models where these gradients are unavailable in closed form, the use of numerical approximations could provide a successful strategy (Calderhead and Sustik, 2012). Results on benchmark datasets demonstrate that CTI outperforms standard TI, but that the difference in performance is reduced when the likelihood function is strongly multi-modal. A natural direction for further research is to explore whether

alternative control variates are better suited to these challenging problems.

CTI clearly inherits the theoretical and methodological challenges that are associated with control variates more generally. In particular ZV control variates are not parametrisation-invariant and it is unclear how to select an optimal variance-minimising parametrisation. Pertinent to CTI in particular, the optimal coefficients $\phi^*(t)$ will vary smoothly with (inverse) temperature t (SFig. 6b), yet the conventional plug-in approach to estimation treats each rung t_i of the temperature ladder independently, leading to rough trajectories (SFig. 6a). It would therefore be interesting to design an information sharing scheme that jointly estimates all coefficients.

The development of low-cost computational approaches to Bayesian model comparison is necessary for the widespread adoption of Bayesian methodology in hypothesis-driven research. The extension of control variate strategies to this important setting offers a promising route towards achieving this goal.

A Quadrature for TI

Implementations of TI employ quadrature to approximate the one dimensional integral in Eqn. 3. Friel and Pettitt (2008) originally employed a simple trapezoidal rule whereby the (inverse) temperature domain $t \in [0, 1]$ was partitioned using $0 = t_0 < t_1 < \dots < t_m = 1$ and the (log-)evidence was approximated by

$$\log(p(\mathbf{y})) \approx \sum_{i=0}^{m-1} \frac{(t_{i+1} - t_i)}{2} [\mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}, t_i} \log(p(\mathbf{y}|\boldsymbol{\theta})) + \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}, t_{i+1}} \log(p(\mathbf{y}|\boldsymbol{\theta}))]. \quad (31)$$

The use of quadrature introduces bias into the resulting estimator. To reduce this quadrature error and thus the estimator bias, Friel *et al.* (2014) proposed the second order correction term

$$\sum_{i=0}^{m-1} \frac{(t_{i+1} - t_i)^2}{12} [\mathbb{V}_{\boldsymbol{\theta}|\mathbf{y}, t_{i+1}} \log(p(\mathbf{y}|\boldsymbol{\theta})) - \mathbb{V}_{\boldsymbol{\theta}|\mathbf{y}, t_i} \log(p(\mathbf{y}|\boldsymbol{\theta}))] \quad (32)$$

that is subtracted from Eqn. 31. Here $\mathbb{V}_{\boldsymbol{\theta}|\mathbf{y},t}g(\boldsymbol{\theta})$ denotes the variance of the function $g(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ has distribution with density $p(\boldsymbol{\theta}|\mathbf{y},t)$.

B Asymptotic unbiasedness

Propositions 1 and 2 of Mira *et al.* (2013) show that a sufficient conditions for asymptotic unbiasedness of ZV control variates, i.e. $\mathbb{E}_\pi[h(\boldsymbol{\theta})] = 0$, is that, in the case where Θ is unbounded, $\lim_{r \nearrow \infty} \int_{\partial B_r} \pi \nabla P \cdot \mathbf{n} d\sigma = 0$ where $B_r \nearrow \Theta$ is a sequence of bounded subsets and \mathbf{n} denotes the vector orthogonal to the boundary ∂B_r . This condition could be difficult to verify directly; below we contribute a sufficient condition $\Theta = \mathbb{R}^d$ that is easily verified. Consider a d -dimensional hypercube $B_r = \{\boldsymbol{\theta} : |\theta_i| \leq r/2\}$ with side length r and surface area $2dr^{d-1}$ and let k be the degree of the polynomial $P(\boldsymbol{\theta})$. Then crude bounds give $\int_{\partial B_r} \pi \nabla P \cdot \mathbf{n} d\sigma \leq \sup_{\boldsymbol{\theta} \in \partial B_r} |\pi(\boldsymbol{\theta}) \nabla P(\boldsymbol{\theta}) \cdot \mathbf{n}(\boldsymbol{\theta})| \times \int_{\partial B_r} d\sigma \leq \left[\sup_{\|\boldsymbol{\theta}\|_1 \geq r} |\pi(\boldsymbol{\theta})| \right] \left[\sup_{\boldsymbol{\theta} \in \partial B_r} \|\nabla P(\boldsymbol{\theta})\|_1 \right] \times 2dr^{d-1}$. Since $\sup_{\boldsymbol{\theta} \in \partial B_r} \|\nabla P(\boldsymbol{\theta})\| = \mathcal{O}(r^{k-1})$ it follows that a sufficient condition for unbiasedness of ZV is

$$\left[\sup_{\|\boldsymbol{\theta}\|_1 \geq r} \pi(\boldsymbol{\theta}) \right] r^{d+k-2} \rightarrow 0 \quad \text{as } r \rightarrow \infty. \quad (33)$$

In practice this requires that the tails of the (unnormalised) density $\pi(\boldsymbol{\theta})$ vanish sufficiently quickly, with faster convergence required when higher degree polynomials are to be used.

C Second degree polynomials

Second degree polynomials can be expressed as $P(\boldsymbol{\theta}) = \mathbf{c}^T \boldsymbol{\theta} + \frac{1}{2} \boldsymbol{\theta}^T \mathbf{B} \boldsymbol{\theta}$ where \mathbf{c} is $d \times 1$ and \mathbf{B} is $d \times d$. This leads to ZV control variates of the form

$$h(\boldsymbol{\theta}) = -\frac{1}{2} \text{tr}(\mathbf{B}) + (\mathbf{c} + \mathbf{B}\boldsymbol{\theta})^T \mathbf{z}(\boldsymbol{\theta}), \quad (34)$$

where \mathbf{c} and \mathbf{B} denote the quadratic polynomial coefficients and $\text{tr}(\mathbf{B})$ is the trace of \mathbf{B} . We assume that \mathbf{B} is symmetric, but this is not required in general. Following Mira *et al.* (2013), it is possible to rearrange the terms on the right hand side of Eqn. 34 into the form $\boldsymbol{\phi}^T \mathbf{w}(\boldsymbol{\theta})$ where the column vectors $\boldsymbol{\phi}$, $\mathbf{w}(\boldsymbol{\theta})$ have $d(d+3)/2$ elements each, and are defined as $\boldsymbol{\phi} := [\mathbf{c}^T \mathbf{d}^T \mathbf{b}^T]^T$, where \mathbf{d} is the diagonal of \mathbf{B} and \mathbf{b} is a column vector with $d(d-1)/2$ elements, whose element in the $(2d-j)(j-1)/2 + (i-j)$ position is the lower diagonal (i, j) -th element of \mathbf{B} , and $\mathbf{w} := [\mathbf{z}^T \mathbf{u}^T \mathbf{v}^T]^T$, where $\mathbf{u} := \boldsymbol{\theta} \circ \mathbf{z} - \frac{1}{2} \mathbf{1}$, with \circ , $\mathbf{1}$ denoting the Hadamard product and the unit vector respectively, while \mathbf{v} is a column vector comprising $d(d-1)/2$ elements, whose element in the $(2d-j)(j-1)/2 + (i-j)$ position equals $\theta_i z_j + \theta_j z_i$, $j \in \{1, 2, \dots, d\}$, $i \in \{2, 3, \dots, d\}$, $j < i$.

The same derivation used to obtain Eqn. 13 can be followed to deduce that optimal coefficients $\boldsymbol{\phi}^*$ in the case of second order polynomials are given by $\boldsymbol{\phi}^* = -\mathbb{V}_\pi^{-1}[\mathbf{w}(\boldsymbol{\theta})] \mathbb{E}_\pi[g(\boldsymbol{\theta}) \mathbf{w}(\boldsymbol{\theta})]$. Similarly the ZV strategy with degree 2 polynomials can be expected to reduce variance when a linear combination of the components of $\mathbf{w}(\boldsymbol{\theta})$ is highly correlated with the target function $g(\boldsymbol{\theta}) = \log p(\mathbf{y}|\boldsymbol{\theta})$.

D Formulae for Bayesian linear regression

D.1 Known precision

The power posterior follows $\boldsymbol{\beta}|\mathbf{y}, t \sim N(\boldsymbol{\mu}(t), \boldsymbol{\Sigma}(t))$ where $\boldsymbol{\mu}(t) = \frac{t}{\sigma^2} \boldsymbol{\Sigma}(t) \mathbf{X}^T \mathbf{y}$, $\boldsymbol{\Sigma}(t)^{-1} = \frac{t}{\sigma^2} \mathbf{X}^T \mathbf{X} + \frac{1}{\zeta^2} \mathbf{I}$, whilst the integrand $\mathbb{E}_{\boldsymbol{\beta}|\mathbf{y}, t}[\log p(\mathbf{y}|\boldsymbol{\beta}, \sigma)]$ has the closed-form expression $-\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\mu}(t))^T (\mathbf{y} - \mathbf{X}\boldsymbol{\mu}(t)) - \frac{1}{2\sigma^2} \text{tr}(\mathbf{X}^T \mathbf{X} \boldsymbol{\Sigma}(t))$ and the model evidence is

$$p(\mathbf{y}) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Omega}|^{1/2}} \exp \left\{ -\frac{1}{2} \mathbf{y}^T \boldsymbol{\Omega}^{-1} \mathbf{y} \right\} \quad (35)$$

where $\boldsymbol{\Omega} = \sigma^2 \mathbf{I} + \zeta^2 \mathbf{X} \mathbf{X}^T$.

D.2 Unknown precision

Using the transformation $\tau \mapsto \eta = \log(\tau)$ we can ensure that the posterior $p(\boldsymbol{\theta}|\mathbf{y}, t, m)$ is defined on \mathbb{R}^d and has exponential tails so that, by Eqn. 33, the unbiasedness condition is satisfied. For Model 1 we have $z_1 = -\frac{1}{2}te^\eta (\sum_i y_i - \alpha - \beta\bar{x}_i) + \frac{1}{2}e^\eta r_0(\alpha - 3000)$, $z_2 = -\frac{1}{2}te^\eta (\sum_i (y_i - \alpha - \beta\bar{x}_i)\bar{x}_i) + \frac{1}{2}e^\eta s_0(\beta - 185)$ and $z_3 = -\frac{nt}{4} + \frac{te^\eta}{4} (\sum_i (y_i - \alpha - \beta\bar{x}_i)^2) - \frac{1+a_0}{2} + \frac{e^\eta}{2} [b_0 + \frac{r_0}{2}(\alpha - 3000)^2 + \frac{s_0}{2}(\beta - 185)^2]$, where the components are ordered with respect to $\boldsymbol{\theta} = (\alpha, \beta, \eta)$.

Write \mathbf{X} for the design matrix with i th row $[1, \bar{x}_i]$. The model evidence, that is the object we wish to estimate, is given by

$$p(\mathbf{y}) = \frac{b_0^{a_0}}{(2\pi)^{n/2}} \sqrt{\frac{|Q_0|}{|Q_n|}} \frac{\Gamma(a_n)}{\Gamma(a_0)} \left\{ b_0 + \frac{1}{2} [\mathbf{y}'\mathbf{y} - \mathbf{B}_n^T \mathbf{Q}_n \mathbf{B}_n + \mathbf{B}_0^T \mathbf{Q}_0 \mathbf{B}_0] \right\}^{-a_n} \quad (36)$$

where $a_n = a_0 + \frac{n}{2}$, $\mathbf{Q}_n = \mathbf{Q}_0 + \mathbf{X}^T \mathbf{X}$ and $\mathbf{B}_n = \mathbf{Q}_n^{-1}(\mathbf{X}^T \mathbf{y} + \mathbf{Q}_0 \mathbf{B}_0)$. Derivations for Model 2 are analogous.

E Formulae for Goodwin Oscillator

The Goodwin oscillator with g species is given by

$$\begin{aligned} \frac{dx_1}{ds} &= \frac{a_1}{1 + a_2 x_g^\rho} - \alpha x_1 \\ \frac{dx_2}{ds} &= k_1 x_1 - \alpha x_2 \\ &\vdots \\ \frac{dx_g}{ds} &= k_{g-1} x_{g-1} - \alpha x_g. \end{aligned} \quad (37)$$

Here x_1 represents the concentration of mRNA for a target gene and x_2 represents its corresponding protein product. Additional variables x_3, \dots, x_g represent intermediate protein species that facilitate a cascade of enzymatic activation that ultimately leads to a negative feedback, via x_g , on the rate at which mRNA is transcribed. The so-

lution $\mathbf{x}(s; \boldsymbol{\theta}, \mathbf{x}_0)$ of this dynamical system depends upon synthesis rate constants a_1, k_1, \dots, k_{g-1} and degradation rate constants a_2, α . The Goodwin oscillator permits oscillatory solutions only when $\rho > 8$. Following Calderhead and Girolami (2009) we therefore set $\rho = 10$ as a fixed parameter. A g -variable Goodwin model as described above therefore has $g + 2$ uncertain parameters $(a_1, a_2, k_1, \dots, k_{g-1}, \alpha)$. The Goodwin oscillator does not permit a closed form solution, meaning that each evaluation of the likelihood function requires the numerical integration of the system in Eqn. 37. Due to the substantive computational challenges associated with model comparison in this setting, we considered only 10 independent runs of population MCMC, each using only $N = 1,000$ iterations.

We consider a realistic setting where only mRNA and protein product are observed, corresponding to $\mathbf{x}_a = [x_1, x_2]$. We assume $\mathbf{x}_0 = [0, \dots, 0]$ and $\sigma = 0.1$ are both known and take sampling times to be $s = 41, \dots, 80$. Parameters were assigned independent $\Gamma(2, 1)$ prior distributions. We generated data using $a_1 = 1, a_2 = 3, k_1 = 2, k_2, \dots, k_{g-1} = 1, \alpha = 0.5$, which produce oscillatory dynamics that do not depend heavily upon initial conditions (SFig. 8).

In practice we work with the log-transformed parameters $\boldsymbol{\theta}$. In particular this allows us to verify that ZV methods are valid, since the tails of $p(\boldsymbol{\theta})$ vanish exponentially quickly. Sensitivities $S_{j,k}^i$, defined in the main text, satisfy

$$\dot{S}_{j,k}^i = \frac{\partial f_k}{\partial \theta_i} + \sum_l \frac{\partial f_k}{\partial x_l} S_{j,l}^i \quad (38)$$

where $\frac{\partial x_k}{\partial \theta_i} = 0$ at $s = 0$. Eqn. 38 provides a route to compute the sensitivities numerically, when they cannot be obtained analytically, by augmenting the state vector of the dynamical system to include the $S_{j,k}^i$.

References

- Andradóttir *et al.* (1993), Variance reduction through smoothing and control variates for Markov Chain simulations. *ACM T. M. Comput. S.* **3**(3):167-189.
- Assaraf, R., and Caffarel, M. (1999), Zero-Variance Principle for Monte Carlo Algorithms. *Phys. Rev. Lett.* **83**(23):4682–4685.
- Behrens, G., Friel, N., and Hurn, M. (2012), Tuning tempered transitions. *Stat. Comput.* **22**(1):65-78.
- Calderhead, B., and Girolami, M. (2009), Estimating Bayes factors via thermodynamic integration and population MCMC. *Comput. Stat. Data An.* **53**(12):4028-4045.
- Calderhead, B., and Girolami, M. (2011), Statistical analysis of nonlinear dynamical systems using differential geometric sampling methods. *Interface Focus* **1**(6):821-835
- Calderhead, B., and Sustik, M. (2012), Sparse Approximate Manifolds for Differential Geometric MCMC. *Adv. Neur. In.* **25**:2888-2896.
- Chen, M. H., Shao, Q. M., and Ibrahim, J. G. (2000), *Monte Carlo methods in Bayesian computation*. Springer New York.
- Chib, S., and Jeliazkov, I. (2001), Marginal likelihood from the Metropolis-Hastings output. *J. Am. Stat. Assoc.* **96**(453):270-281.
- Corduneanu, A., and Bishop, C. M. (2001), Variational Bayesian model selection for mixture distributions. *Proceedings of the 8th International Conference on Artificial intelligence and Statistics* :27-34.
- Del Moral, P., Doucet, A., and Jasra, A. (2006), Sequential monte carlo samplers. *J. R. Statist. Soc. B* **68**(3):411-436.

- Dellaportas, P., and Kontoyiannis, I. (2012), Control variates for estimation based on reversible Markov chain Monte Carlo samplers. *J. R. Statist. Soc. B* **74**(1):133-161
- Didelot *et al.* (2011), Likelihood-free estimation of model evidence. *Bayesian Analysis* **6**(1):49-76.
- Frenkel, D., and Smit, B. (2002), *Understanding Molecular Simulation: From Algorithms to Applications (2nd Edition)*. Academic Press.
- Friel, N., and Pettitt, A. N. (2008), Marginal likelihood estimation via power posteriors. *J. R. Statist. Soc. B* **70**(3):589-607.
- Friel, N., and Wyse, J. (2012), Estimating the statistical evidence – a review. *Stat. Neerl.* **66**:288-308.
- Friel, N., Hurn, M. A., and Wyse, J. (2014), Improving power posterior estimation of statistical evidence. *Stat. Comp.*, in press.
- Gelfand, A. E., and Dey, D. K. (1994), Bayesian model choice: asymptotics and exact calculations. *J. R. Statist. Soc. B* **56**(3):501-514.
- Gelman, A., and Meng, X.-L., (1998), Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Stat. Sci.* **13**(2):163-185.
- Geyer, C. J., and Thompson, E. A. (1995), Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Am. Stat. Assoc.* **90**(431):909-920.
- Glasserman, P. (2004), *Monte Carlo Methods in Financial Engineering*. Springer-Verlag New York.
- Girolami, M., and Calderhead, B. (2011), Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Statist. Soc. B* **73**(2):1-37.

- Goodwin, B. (1965), Oscillatory behavior in enzymatic control processes. *Adv. Enzyme Regul.* **3**:425-438.
- Green, P., and Han, X. (1992), Metropolis methods, Gaussian proposals, and antithetic variables. *Lecture Notes in Statistics, Stochastic Methods and Algorithms in Image Analysis* **74**:142-164.
- Green, P. J. (1995), Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**(4):711-732.
- Hammer, H., and Tjelmeland, H. (2008), Control variates for the Metropolis-Hastings algorithm. *Scand. J. Stat.* **35**(3):400-414.
- Jasra, A., Stephens, D., and Holmes, C. (2007), On population-based simulation for static inference. *Stat. Comput.* **17**(3):263-279.
- Kass, R. E., and Raftery, A. E. (1995), Bayes factors. *J. Am. Stat. Assoc.* **90**(430):773-795.
- Locke, J., Millar, A., and Turner, M. (2005), Modelling genetic networks with noisy and varied experimental data: the circadian clock in arabidopsis thaliana. *J. Theor. Biol.* **234**(3):383-393.
- Marin, J. M., and Robert, C. P. (2011), Importance sampling methods for Bayesian discrimination between embedded models. *Frontiers of Statistical Decision Making and Bayesian Analysis*, Springer New York.
- Marinari, E., and Parisi, G. (1992), Simulated tempering: a new Monte Carlo scheme. *Europhys. Lett.* **19**(6):451.
- Miasojedow, B., Moulines, E., and Vihola, M. (2012), Adaptive Parallel Tempering Algorithm. arXiv 1205.1076.

- Mira, A., Tenconi, P., and Bressanini, D. (2003), Variance reduction for MCMC. Technical Report 2003/29, Università degli Studi dell' Insubria, Italy.
- Mira, A., Solgi, R., and Imparato, D. (2013), Zero Variance Markov Chain Monte Carlo for Bayesian Estimators, *Stat. Comput.* **23**(5):653-662..
- Neal, R. (1996), Sampling from multimodal distributions using tempered transitions. *Stat. Comput.* **6**(4):353-366.
- Neal, R. M. (2001), Annealed importance sampling. *Stat. Comput.* **11**(2):125-139.
- Ogata, Y. (1989), A Monte Carlo method for high dimensional integration. *Numer. Math.* **55**(2):137-157.
- Papamarkou, T., Mira, A., and Girolami, M. (2014), Zero Variance Differential Geometric Markov Chain Monte Carlo Algorithms. *Bayesian Analysis*, in press.
- Philippe, A., and Robert, C. (2001), Riemann sums for MCMC estimation and convergence monitoring. *Stat. Comput.* **11**(2):103-115.
- Robert, C., and Casella, G. (2004), *Monte Carlo Statistical Methods. (2nd ed.)* Springer-Verlag New York.
- Rubinstein, R. Y., and Marcus, R. (1985), Efficiency of Multivariate Control Variates in Monte Carlo Simulation. *Oper. Res.* **33**(3):661-677.
- Skilling, J. (2006), Nested sampling for general Bayesian computation. *Bayesian Analysis* **1**(4):833-859.
- Toni *et al.* (2009), Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface* **6**(31):187-202.
- Torrie, G. M., and Valleau, J. P. (1977), Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **23**(2):187-199.

- Vyshemirsky, V., and Girolami, M. A. (2008), Bayesian ranking of biochemical system models. *Bioinformatics* **24**(6):833-839.
- Williams, E. (1959), *Regression Analysis*. Wiley New York.
- Zhou, Y., Johansen, A. M., and Aston, J. A. D. (2013), Towards Automatic Model Comparison an Adaptive Sequential Monte Carlo Approach. *CRiSM Technical Report, University of Warwick*, 13-04.

Supplement

Proof of exactness

In this section we prove that CTI (degree 2) is exact (up to quadrature error) for the Bayesian linear regression model with known precision.

We have from Eqn. 15 that the minimum variance ratio is given by

$$R = 1 - \max_{\phi} \text{Corr}_{\beta|y,t}[g(\beta), \phi^T \mathbf{w}(\beta)]. \quad (39)$$

Plugging in the expression of Eqn. 34 for $\mathbf{w}(\beta)$ we obtain

$$R = 1 - \max_{\mathbf{B}, \mathbf{c}} \text{Corr}_{\beta|y,t}[g(\beta), -\frac{1}{2} \text{tr}(\mathbf{B}) + (\mathbf{c} + \mathbf{B}\beta)^T \mathbf{z}(\beta)] \quad (40)$$

where the maximum is taken over all symmetric matrices \mathbf{B} and real vectors \mathbf{c} .

Write $\stackrel{+C}{=}$ whenever two quantities are equal up to an additive constant not depending upon β ; since $\text{Corr}_{\beta|y,t}[W, X] = \text{Corr}_{\beta|y,t}[Y, Z]$ whenever $W \stackrel{+C}{=} Y$ and $X \stackrel{+C}{=} Z$, we need only work up to this equivalence. We now claim that $\mathbf{z}(\beta)$ can be replaced with any transformation $\mathbf{z} \mapsto \mathbf{f} + \mathbf{E}\mathbf{z}$ in Eqn. 40, where we require that \mathbf{E} is symmetric and invertible. Indeed

$$(\mathbf{c} + \mathbf{B}\beta)^T (\mathbf{f} + \mathbf{E}\mathbf{z}(\beta)) \stackrel{+C}{=} (\mathbf{c}' + \mathbf{B}'\beta)^T \mathbf{z}(\beta) + \mathbf{f}^T \mathbf{E}\beta \quad (41)$$

where $\mathbf{c}' = \mathbf{E}^T \mathbf{c}$, $\mathbf{B}' = \mathbf{E}^T \mathbf{B}$ (which is symmetric). Moreover, from the definition of the control variates (Eqn. 20) we have that $\beta = 2\boldsymbol{\Sigma}(t)[\mathbf{z}(\beta) + \frac{t}{2\sigma^2} \mathbf{X}^T \mathbf{y}]$ and hence

$$\mathbf{f}^T \mathbf{E}\beta \stackrel{+C}{=} (\mathbf{c}'')^T \mathbf{z}(\beta) \quad (42)$$

where $\mathbf{c}'' = 2\mathbf{f}^T \mathbf{E}\Sigma(t)$. Combining Eqs. 41 and 42 we have that

$$(\mathbf{c} + \mathbf{B}\beta)^T (\mathbf{f} + \mathbf{E}\mathbf{z}(\beta)) \stackrel{+C}{=} (\mathbf{c}''' + \mathbf{B}'\beta)^T \mathbf{z}(\beta) \quad (43)$$

where $\mathbf{c}''' = \mathbf{c}'' + \mathbf{c}''$. Recalling that correlation is invariant to the addition of constant terms, we have shown that

$$R \leq 1 - \max_{\mathbf{B}, \mathbf{c}} \text{Corr}_{\beta|\mathbf{y}, t}[g(\beta), -\frac{1}{2}\text{tr}(\mathbf{B}) + (\mathbf{c} + \mathbf{B}\beta)^T (\mathbf{f} + \mathbf{E}\mathbf{z}(\beta))]. \quad (44)$$

In fact this equation is an equality, since the affine transformation is invertible and hence we can apply the same argument using the inverse transform.

Now $g(\beta) \stackrel{+C}{=} (\beta - \mathbf{m})^T \mathbf{S}^{-1}(\beta - \mathbf{m})$ where $\mathbf{S} = (\mathbf{X}^T \mathbf{X} / \sigma^2)^{-1}$, $\mathbf{m} = \mathbf{S} \mathbf{X}^T \mathbf{y} / \sigma^2$. Taking the specific choices $\mathbf{B} = \mathbf{S}^{-1}$ (which is symmetric), $\mathbf{c} = -\mathbf{S}^{-1} \mathbf{m}$, $\mathbf{f} = \frac{t}{\sigma^2} \Sigma(t) \mathbf{X}^T \mathbf{y} - \mathbf{m}$ and $\mathbf{E} = 2\Sigma(t)$ (which is symmetric and invertible) we have

$$R \leq 1 - \text{Corr}_{\beta|\mathbf{y}, t}[(\beta - \mathbf{m})^T \mathbf{S}^{-1}(\beta - \mathbf{m}), (\beta - \mathbf{m})^T \mathbf{S}^{-1}(\beta - \mathbf{m})] = 1 - 1 = 0 \quad (45)$$

which demonstrates that $R = 0$ and CTI (degree 2) is exact.

Manifold Metropolis-Adjusted Langevin Algorithm

mMALA is a differential geometric MCMC scheme that, for power posteriors, requires that we have access to the metric tensor

$$\mathbf{G}(\boldsymbol{\theta}|t) = -\mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}} \frac{\partial^2}{\partial \boldsymbol{\theta}^2} \log p(\mathbf{y}, \boldsymbol{\theta}|t). \quad (46)$$

At current state $\boldsymbol{\theta}_n^{(i)}$ and for (inverse) temperature t_i the “simplified” mMALA proposal follows from a discretised Langevin diffusion

$$\boldsymbol{\theta}^* | \boldsymbol{\theta}_n^{(i)}, \mathbf{y}, t_i \sim N \left(\boldsymbol{\theta}_n^{(i)} + \frac{\epsilon^2}{2} \mathbf{G}^{-1}(\boldsymbol{\theta}_n^{(i)} | \mathbf{y}, t_i) \nabla_{\boldsymbol{\theta}} \log[p(\mathbf{y}, \boldsymbol{\theta}_n^{(i)} | t_i)], \epsilon^2 \mathbf{G}^{-1}(\boldsymbol{\theta}_n^{(i)} | \mathbf{y}, t_i) \right) \quad (47)$$

that assumes constant curvature of the manifold. The proposal $\boldsymbol{\theta}^*$ is then accepted as the next state $\boldsymbol{\theta}_{n+1}^{(i)}$ according to the Metropolis-Hastings ratio (else $\boldsymbol{\theta}_{n+1}^{(i)} = \boldsymbol{\theta}_n^{(i)}$). For all applications in this paper we discarded the first 10% of samples as burn-in and then retained the remaining N samples for use.

The metric tensors for each of the applications considered in the Main Text are provided below:

Bayesian linear regression, known precision.

$$\mathbf{G}(\boldsymbol{\beta}|t) = \frac{t}{\sigma^2} \mathbf{X}^T \mathbf{X} - \frac{1}{\zeta^2} \mathbf{I}_{d \times d} \quad (48)$$

Bayesian linear regression, unknown precision (Radiata Pine).

$$\mathbf{G}(\boldsymbol{\theta}|t) = \begin{bmatrix} e^\eta(nt + r_0) & 0 & e^\eta r_0(\alpha - 3000) \\ 0 & e^\eta (s_0 + t \sum_i \bar{x}_i^2) & e^\eta s_0(\beta - 185) \\ e^\eta r_0(\alpha - 3000) & e^\eta s_0(\beta - 185) & \frac{tn}{2} + e^\eta (b_0 + \frac{r_0}{2}(\alpha - 3000)^2 + \frac{s_0}{2}(\beta - 185)^2) \end{bmatrix} \quad (49)$$

Bayesian logistic regression (Pima Indians).

$$G_{j,k}(\boldsymbol{\beta}|t) = -t \sum_i p_i(1 - p_i)x_{i,j}x_{i,k} + \tau\delta_{j,k} \quad (50)$$

Bayesian inference for nonlinear ODEs (Goodwin Oscillator).

$$G_{i,l}(\boldsymbol{\theta}|t) = \delta_{i,l} \exp(\theta_i) + \frac{t}{\sigma^2} \sum_j [\mathbf{S}_{j,\bullet}^i][\mathbf{S}_{j,\bullet}^l]^T \quad (51)$$

The controlled (equilibrated) annealed importance sampler

Annealed importance sampling (AIS) was proposed by Neal (2001) as an extension of bridge sampling that improves mixing in parameter space by introducing multiple intermediate densities. In brief, AIS proceeds by producing samples $\boldsymbol{\theta}^{(0)}, \dots, \boldsymbol{\theta}^{m-1}$ as follows: $\boldsymbol{\theta}^{(0)} \sim p(\boldsymbol{\theta})$. Then $\boldsymbol{\theta}^{(j)} \sim T_j(\boldsymbol{\theta}^{(j-1)})$ in sequence for $j = 1, \dots, m-1$

where T_j is a Markov transition kernel that targets the distribution $\boldsymbol{\theta}|\mathbf{y}, t = t_j$. Let $f(\boldsymbol{\theta}|\mathbf{y}, t) = p(\boldsymbol{\theta}|\mathbf{y})^t p(\boldsymbol{\theta})$ so that $p(\boldsymbol{\theta}|\mathbf{y}, t) = f(\boldsymbol{\theta}|\mathbf{y}, t)/\mathcal{Z}_t(\mathbf{y})$. Define

$$w = \frac{f(\boldsymbol{\theta}^{(0)}|\mathbf{y}, t_1)}{f(\boldsymbol{\theta}^{(0)}|\mathbf{y}, t_0)} \cdot \frac{f(\boldsymbol{\theta}^{(1)}|\mathbf{y}, t_2)}{f(\boldsymbol{\theta}^{(1)}|\mathbf{y}, t_1)} \cdots \frac{f(\boldsymbol{\theta}^{(m-1)}|\mathbf{y}, t_m)}{f(\boldsymbol{\theta}^{(m-1)}|\mathbf{y}, t_{m-1})}. \quad (52)$$

Then it is shown in Neal (2001) that

$$\mathbb{E}_{(\boldsymbol{\theta}^{(0)}, \dots, \boldsymbol{\theta}^{(m-1)}) \sim G}[w] = \frac{\mathcal{Z}_1}{\mathcal{Z}_0} \cdot \frac{\mathcal{Z}_2}{\mathcal{Z}_1} \cdots \frac{\mathcal{Z}_m}{\mathcal{Z}_{m-1}} = \frac{\mathcal{Z}_m}{\mathcal{Z}_0} = p(\mathbf{y}) \quad (53)$$

where the expectation is over the generative process G described above. Note that this is precisely m versions of bridge sampling, each targeting one of the ratios in the above equation.

AIS is a non-equilibrium estimator, in the sense that the marginal distribution of $\boldsymbol{\theta}^{(i)}$ need not be the same as the distribution $\boldsymbol{\theta}|\mathbf{y}, t_i$, and is therefore not directly amenable to ZV control variates. In order to transform AIS into an equilibrium estimator we need to consider jointly sampling all the $\boldsymbol{\theta}^{(i)}$. Specifically, we exploit the fact that

$$\mathbb{E}_{(\boldsymbol{\theta}^{(0)}, \dots, \boldsymbol{\theta}^{(m-1)}) \sim G}[w] = \mathbb{E}_{\substack{\boldsymbol{\theta}^{(i)}|\mathbf{y}, t_i \\ 0 \leq i \leq m-1}}[w]. \quad (54)$$

Estimation in the equilibrated AIS therefore requires a collection of samples $\boldsymbol{\theta}^{(j)} \sim \boldsymbol{\theta}|\mathbf{y}, t_j$ that can be obtained using (converged) MCMC. In this paper we generated these samples using population MCMC (Jasra *et al.*, 2007); for fair comparison we used the same samples that were the basis for TI experiments.

Rewriting w as in Vyshemirsky and Girolami (2008) we obtain

$$p(\mathbf{y}) = \mathbb{E}_{\substack{\boldsymbol{\theta}^{(i)}|\mathbf{y}, t_i \\ 0 \leq i \leq m-1}} \left[\exp \left(\sum_{i=0}^{m-1} (t_{i+1} - t_i) \log(p(\mathbf{y}|\boldsymbol{\theta}^{(i)})) \right) \right]. \quad (55)$$

Since a Monte Carlo estimate based on Eqn. 55 will be unbiased, we need simply choose the temperature ladder sufficiently fine that our acceptance rates indicate good

mixing. In experiments below, for fairness of comparison, the same temperature ladder was used for (C)AIS as for (C)TL. This controls the amount of information present in the samples $\boldsymbol{\theta}_n^{(i)}$ and allows the samples from the same run of population MCMC to be used for all estimators.

The Monte Carlo expectation for equilibrated AIS is taken over all $\boldsymbol{\theta}^{(0:m-1)} = \{\boldsymbol{\theta}^{(i)}\}_{i=0}^m$ simultaneously; we therefore base ZV control variates on

$$\mathbf{z}(\boldsymbol{\theta}^{(0:m-1)}|\mathbf{y}, t_{0:m-1}) = -\frac{1}{2}\nabla_{\boldsymbol{\theta}^{(0:m-1)}} \log \left[\prod_{i=0}^{m-1} p(\boldsymbol{\theta}^{(i)}|\mathbf{y}, t_i) \right] \quad (56)$$

so that $\mathbf{z}(\boldsymbol{\theta}^{(0:m-1)}|\mathbf{y}, t_{0:m-1})$ has a block structure whose components are given by Eqn. 7. Then ZV control variates are given by

$$\begin{aligned} h(\boldsymbol{\theta}^{(0:m-1)}|\mathbf{y}, t_{0:m-1}) &= -\frac{1}{2}\Delta_{\boldsymbol{\theta}^{(0:m-1)}}[P(\boldsymbol{\theta}^{(0:m-1)}|\boldsymbol{\phi}(\mathbf{y}, t_{0:m-1}))] \\ &\quad + \nabla_{\boldsymbol{\theta}^{(0:m-1)}}[P(\boldsymbol{\theta}^{(0:m-1)}|\boldsymbol{\phi}(\mathbf{y}, t_{0:m-1}))] \cdot \mathbf{z}(\boldsymbol{\theta}^{(0:m-1)}|\mathbf{y}, t_{0:m-1}). \end{aligned} \quad (57)$$

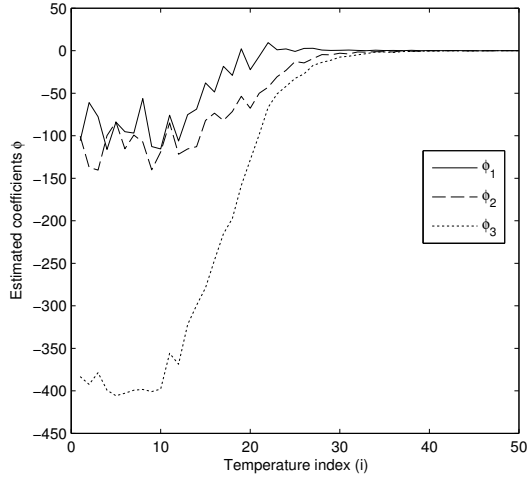
The CAIS estimator is defined by the identity

$$p(\mathbf{y}) = \mathbb{E}_{\substack{\boldsymbol{\theta}^{(i)}|\mathbf{y}, t_i \\ 0 \leq i \leq m-1}} \left[\exp \left(\sum_{i=0}^{m-1} (t_{i+1} - t_i) \log(p(\mathbf{y}|\boldsymbol{\theta}^{(i)})) \right) + h(\boldsymbol{\theta}^{(0:m-1)}|\mathbf{y}, t_{0:m-1}) \right]. \quad (58)$$

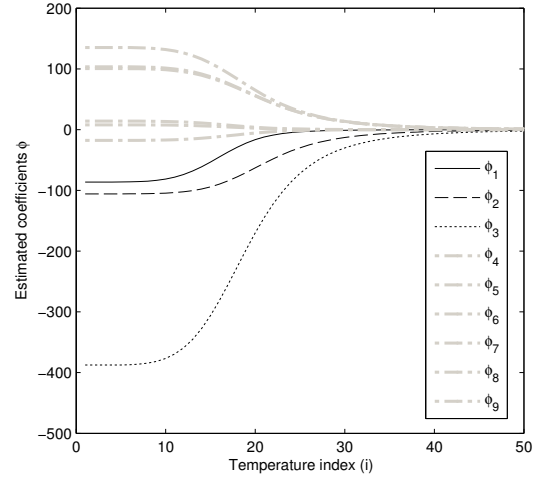
When coefficients $\boldsymbol{\phi}(\mathbf{y}, t_{0:m-1})$ are chosen optimally, the Monte Carlo estimator of Eqn. 58 will have variance that is, in the worst case, no larger than the variance of the standard AIS estimator. In practice, polynomial coefficients are estimated using the plug-in approach of Eqn. 17, taking $g(\boldsymbol{\theta}) = \exp \left(\sum_{i=0}^{m-1} (t_{i+1} - t_i) \log(p(\mathbf{y}|\boldsymbol{\theta}^{(i)})) \right)$.

As discussed in the main text, the plug-in approach typically fails due to the high-dimensionality of the covariance matrices that must be estimated. In addition, implementation of CAIS is complicated due to the requirement that the integrand of Eqn. 58 must remain positive; this further detracts from the suitability of CAIS.

Additional figures



(a) Degree 1



(b) Degree 2

Figure 6: Estimated polynomial coefficients $\phi^*(t_i)$ for ZV control variates. (a) Degree 1 polynomials. (b) Degree 2 polynomials. [Here we show one particular realisation corresponding to one run of population MCMC. It can be seen that, for degree 2 polynomials, the plug-in estimate for coefficients is deterministic. The x-axis records the index i corresponding to (inverse) temperature $t_i = (i/50)^5$.]

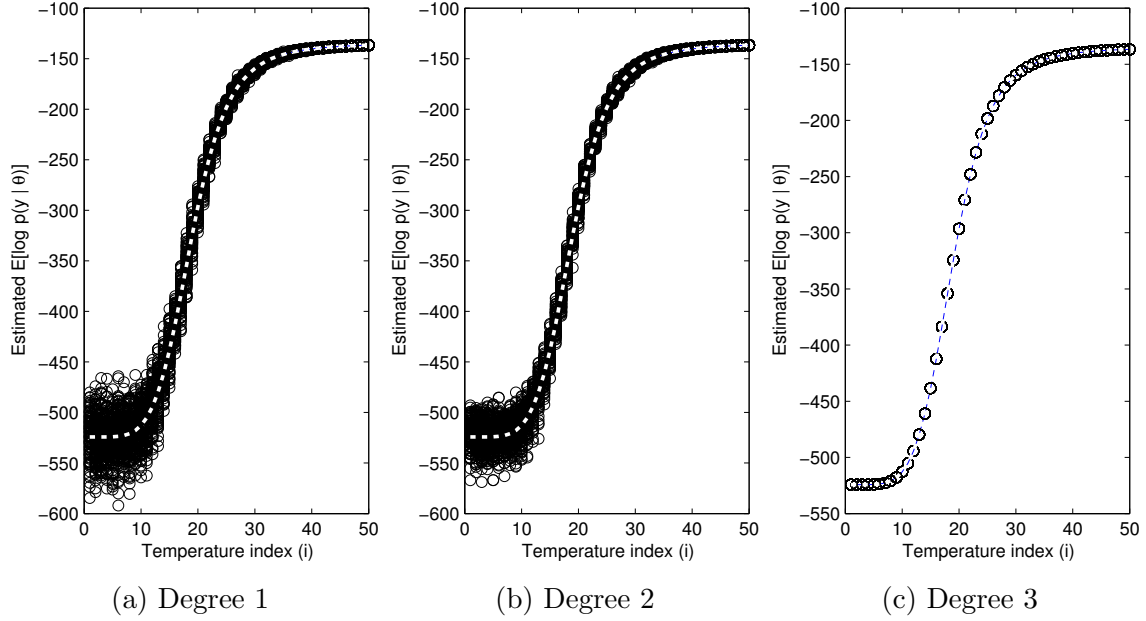


Figure 7: Estimates for the integrand $\mathbb{E}_{\beta|y,t}[\log p(\mathbf{y}|\boldsymbol{\theta})]$, based on 100 independent runs of population MCMC with $N = 1000$ samples and a quintic temperature ladder $t_i = (i/50)^5$. The dashed blue/white curve represents the true value of the integrand. [Here we consider polynomial trial functions $P(\boldsymbol{\theta})$ of (a) degree 0 (i.e. standard TI), (b) degree 1 and (c) degree 2. The x-axis records the index i corresponding to (inverse) temperature $t_i = (i/50)^5$.]

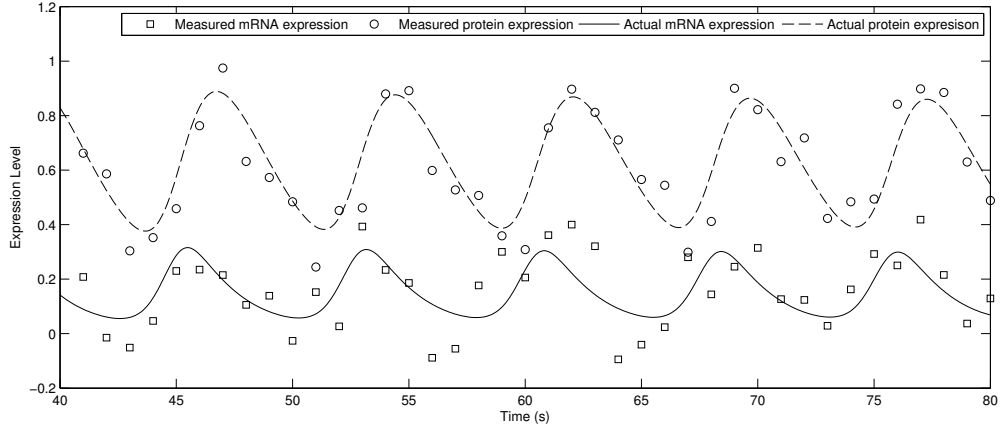


Figure 8: Nonlinear ODEs: Data generated from the Goodwin oscillator based on $g = 3$ species demonstrates characteristic oscillatory behaviour.

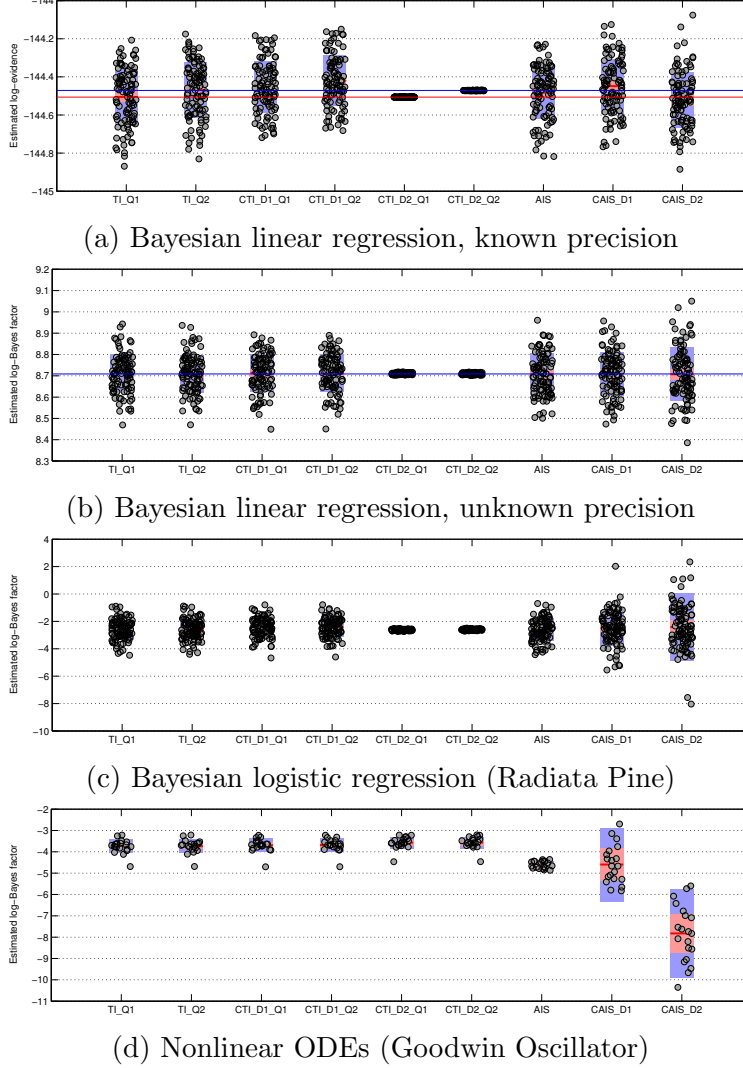


Figure 9: Estimates for evidence/Bayes factors. (a) Bayesian linear regression, known precision: Estimates of log-evidence, based on 100 independent runs of population MCMC with $N = 1000$ samples. The blue line shows the true log-evidence, whereas the red line displays the biased form of the log-evidence when first order quadrature error is taken into account. (b) Radiata pine: Estimates of the log-Bayes factor of Model 2 in favour of Model 1, based on 100 independent runs of population MCMC with $N = 1000$ samples. The blue line shows the true log-Bayes factor, which is $B_{12} = 8.7086$. (c) Pima Indians: Estimates of the log-Bayes factor of Model 2 in favour of Model 1, based on 100 independent runs of population MCMC with $N = 1000$ samples. (d) Goodwin oscillator: Estimates of the log-Bayes factor of Model 2 in favour of Model 1, based on 10 independent runs of population MCMC with $N = 1000$ samples. [TI = thermodynamic integration, CTI = controlled TI, AIS = annealed importance sampling, CAIS = controlled AIS, D1 = degree 1 polynomials, D2 = degree 2 polynomials, Q1 = first order quadrature, Q2 = second order quadrature. Red error regions are used to display 95% confidence intervals for the sample mean over all estimates, and blue error regions display the inter-quartile range for the estimates.]